Book Review

Data Orchestration in Deep Learning Accelerators

Tushar Krishna
Hyoukjun Kwon
Angshuman Parashar
Michael Pellauer
Ananda Samajdar

Synthesis Lectures on Computer Architecture
Morgan & Claypool Publishers

www.morganclaypool.com

ISBN: 9781681738697 paperback
9781681738703 ebook
9781681738710 hardcover

Understanding the basic computer architecture is essential to learn computing perfectly. Deep Learning enables the learning of computer architecture within the specialized accelerators for Deep Learning. The purpose of this book in the words of the authors is that it helps to know the data movement within an accelerator for performance and energy efficiency.

In the chapter 1 on Introduction to Data Orchestration, the authors presented the elementary description of the deep neural networks which is fundamentally a computer mechanism. The training and inference in the deep learning are compute and memory intensive processes. The custom accelerators for DNN inference is performed in the end-user devices and it is essential to know its process in the data orchestration. To comprehend the understanding, the authors described the DNN architecture and DNN models in the base chapter.

The second chapter on Dataflow and Data Reuse has outlined the data reuse opportunities in common operations in the deep neural networks. As the DNN activities involve large number of computations, for which the algorithmic reuse can formalize the choices of the computations for selection of dataflow and mappings. This unit in nutshell described the dataflow and mapping and how mapping and dataflow affect data reuse opportunities with the help of large number of illustrations.

In the third chapter on Buffer hierarchies, the key component of data orchestration is explained. The framework essential for understanding the key option is presented with notes on trade-off between design effort and cross-project reuse. The taxonomy of the buffer hierarchies with the support of algorithms makes this chapter as significant. The buffer storage and implementation is supported with a good number of illustrations.

To explore the network-on-chop architectures in the DNN accelerators, the on-chip data movement characteristics is explained in the fourth chapter on 'Networks on chip'. To do this, the readers can first understand the various traffic movement patterns in the typical DNN accelerators. Later the NoC design is explained with discussion of traffic movement patterns within DNN accelerators.

In the chapter 5 on 'Architecting a DNN Accelerator, the authors discussed the flow employed while designing a DNN accelerator. Then authors have described the decisions one can make when architecting for specific user cases. The sixth chapter on 'Modelling Accelerator Design Space', the authors first explained the mapping step for design-space exploration for DNN accelerators. They have further investigated how microarchitecture models for DNN accelerators can be constructed.

In the next chapter on 'Orchestrating Compressed Sparse Data', the authors have provided background about the sparsity in DNN and discussed the state-of-the art dataflow styles on compressed sparse data.

The last chapter, Conclusion the authors have outlined the possibilities for further research in the data orchestration which is a unique kind of exercise.

As like other Morgan books, an extensive bibliography supports the reading of book.

Hathairat Ketmaneechairat
King Mongkut's University of Technology, North Bangkok
Thailand